Table 9 Summary of RL-based decentralized, partially decentralized, and centralized routing models

Ref.	Technique (selection)	Application (network)	Dataset	Features[a]	Action set	Evaluation: Settings[a]	Evaluation: Improvement[b]
AdaR [461]	Partially decentralized LSPI (*ϵ-greedy*)	Unicast routing (*WSN*)	Simulations ·400 sensors ·20 data sources ·1 sink	State: $\mathcal{N}_i$ Reward: function of · node load · residual energy · hop cost to sink · link reliability	Next-hop nodes to destination	· $S = \#nodes$ · $A = \#neighbors$	Compared to Q-learning: · Faster convergence (by 40 episodes) · Less sensitive to initial parameters
FROMS [151]	Q-learning (*variant of ϵ-greedy*)	Multicast routing (*WSN*)	Omnet++ Mobility Framework with 50 random topologies ·50 nodes ·5 sources ·45 sinks	State: $(\mathcal{N}_j^k, D_k)$ Reward: function of hop cost	$\{a_1 \cdots a_m\}$ $a_k = (\mathcal{N}_j^k, D_k)$ N_j^k = next hop along the path to sink D_k	· $S = \#nodes$ · $A = \#neighbors$	Compared to directed diffusion: · up to 5× higher delivery rate · ≈ 20% lower overhead
Q-PR [24]	Variant of Q-learning (*ϵ-greedy*)	Localization-aware routing to achieve a trade-off between packet delivery rate, ETX, and network lifetime (*WSN*)	Simulations ·50 different topologies ·100 nodes	State: $\mathcal{N}_i$ Reward: function of · distance($\mathcal{N}_i$,$\mathcal{N}_j$) · distance($\mathcal{N}_j$,d) · energy at $\mathcal{N}_j$ · ETX · $\mathcal{N}_j$'s neighbors for any neighbor $\mathcal{N}_j$ and destination	Next-hop nodes to destination	· $S = \#nodes$ · $A = \#neighbors$	Delivery rate: · 25% more than GPSR Network lifetime · 3× more than GPSR · 4× more than EFE
Ref.	Technique (selection)	Application (network)	Dataset	Features[a]	Action set	Evaluation: Settings[a]	Improvement[b]
Xia et al. [482]	DRQ-learning (*greedy*)	Spectrum-aware routing (*CRN*)	OMNET++ simulations · stationary multi-hop CRN · 10 nodes · 2 PUs	State: $\mathcal{N}_i$ Reward: # available channels between current node and next-hop node	Next-hop nodes to destination	· $S = \#nodes$ · $A = \#neighbors$	Compared to Q-routing: ·50% faster at lower activity level Compared to Q-routing and SP-routing: · lower converged end-to-end delay
QELAR [197]	Model-based Q-learning (*greedy*)	Distributed energy-efficient routing (*underwater WSN*)	Simulations (ns-2) ·250 sensors in $500^3 m^3$ space ·100m transmission range · fixed source/sink ·1m/s maximum speed for intermediate nodes	State: $\mathcal{N}_i$ Reward: function of the residual energy of the node receiving the packet and the energy distribution among its neighbor nodes.	Next-hop nodes to destination ∪ packet withdrawal	· $S = \#nodes$ · $A = 1 + \#neighbors$	Compared to Q-learning: · Faster convergence (40 episodes less) · Less sensitive to initial parameters

Table 9 Summary of RL-based decentralized, partially decentralized, and centralized routing models *(Continued)*

Lin et al. [277]	$n-$step TD *(greedy)*	Delay-sensitive application routing *(multi-hop wireless ad hoc networks)*	Simulations 2 users transmitting video sequences to the same destination node ·3 $\sim$ 4-hops wireless network	State: current channel states and queue sizes at the nodes in each hop Reward: goodput at destination	Next-hop nodes to destination	· $S = n_q^N \times n_c^H$ · $A = (N_h^2)^{H-1} \times N_h$ N = *#nodes* N_h = *#nodes at hop h* H = *#hops* n_q = *#queue states* n_c = *#channel states*	Complexity$\approx 2 \times 10^8$ for the 3$-$hop network With 95% less information exchanges · $\sim$ 10% higher PSNR · slightly slower convergence (+1 $\sim$ 2*sec*)
d-AdaptOR [59]	Q-learning with adaptive learning rate *($\epsilon-$greedy)*	Opportunistic routing *(multi-hop wireless ad hoc networks)*	Simulations on QualNet with 36 randomly placed wireless nodes in a 150m × 150m	State: $\mathcal{N}_i$ Reward: · fixed negative transmission cost is receiver is not the destination · fixed positive reward if receiver is the destination · 0 if packet is withdrawn	Next-hop nodes to destination ∪ packet withdrawal	· S = *#nodes* · A = 1 + *#neighbors*	After convergence ($\approx$ 300*sec*) · ETX comparable to a topology-aware routing algorithm · > 30% improvement over greedy-SR, greedy ExOR and SRCR with a single flow · Improvement decreases with # flows
QAR [276]	Centralized SARSA *(ϵ-greedy)*	QoS-aware adaptive routing *(SDN)*	Sprint GIP network trace-driven simulations [418] · 25 switches, 53 links	State: $\mathcal{N}_i$ Reward: function of delay, loss, throughput	Next-hop nodes to destination	· S = *#nodes* · A = *#neighbors*	Compared to Q-learning with QoS-awareness: · Faster convergence time (20 episodes less)

[a] $\mathcal{N}_i$: node i; D_k: sink k; S: number of state variables; A: number of possible actions per state; #: number of
[b] Average values. Results vary according to experimental settings.